Approximating the Rao's distance between negative binomial distributions. Application to counts of marine organisms

Claude Manté, Saikou Oumar Kidé

Approximating the Rao's distance between negative binomial distributions. Application to counts of marine organisms.

Claude Manté, *Aix-Marseille Université, Université du Sud Toulon-Var, CNRS/INSU, IRD, MIO, UM 110, F13288 Marseille Cedex 09, France*, claude.mante@mio.osupytheas.fr
Saikou Oumar Kidé, *Institut Mauritanien de Recherches Océanographiques et des Pêches, Laboratoire de Biologie et Ecologie des Organismes Aquatiques- BP 22 Nouadhibou Mauritania*, saikoukide@gmail.com

Abstract. While the negative binomial distribution is widely used to model catches of animals, it is noteworthy that the parametric approach is ill-suited from an exploratory point of view. Indeed, the "visual" distance between parameters of several distributions is misleading, since on the one hand it depends on the chosen parametrization and on the other hand these parameters are not commensurable (*i. e.* they measure quite different characteristics). Consequently, we settle the topic of comparing abundance distributions in a well-suited framework: the Riemannian manifold $NB(D_{\mathcal{R}})$ of negative binomial distributions, equipped with the Fisher-Rao metrics. It is then possible to compute an intrinsic distance between species. We focus on computational issues encountered in computing this distance between marine species.

Keywords. Information geometry, abundance distributions, geodesic, cut point

1 Introduction

The statistical analysis of counts of living organisms brings information about the collective behavior of species (schooling, habitat preference, *etc*), possibly associated with their biological characteristics (growth rate, reproductive power, survival rate, *etc*). This task can be implemented in an exploratory setting (see for instance [8, 7] and the references therein), but parametric distributions are also widely used for modeling populations abundance. Thus, the negative binomial (NB) distribution is commonly used to model catches of animals [2, 10, 12, 9].

This distribution is especially relevant for this purpose, because [9]:

1. it arises as a Gamma-Poisson mixture, whose parameters depend on the more or less aggregative behavior of the species, and on the efficiency of the trawl for catching it

2. it arises as the limit distribution of the Kendall's [6] birth-and-death model; in this setting, the parameters depend on the demography of the species (reproductive power, mortality, immigration rate)

3. in addition, it is a natural model for collections (of animals, for instance).

But it is noteworthy that the parametric approach is ill-suited from an exploratory point of view: the "visual" distance between parameters of several NB distributions is misleading, because on the one hand it depends on the chosen parametrization and, on the other hand, these parameters are not commensurable in general (they are associated with completely different characteristics of the species, in the setting of different statistical models). Considering the Riemannian manifold $NB(D_{\mathcal{R}})$ of negative binomial distributions (NB) equipped with the Fisher-Rao metrics, we can compute intrinsic distances between species, on the basis of their counts. Then, the "visual" distance between species approximated through Multidimensional Scaling of the table of Rao's distances (for instance) is a sound dissimilarity measure between species.

2 Notations

Consider a Riemannian manifold $\mathfrak{M}$, and a parametric curve $\alpha : [a, b] \to \mathfrak{M}$; its first derivative with respect to "time" will be denoted $\dot{\alpha}$. A geodesic curve γ connecting two points p and q of $\mathfrak{M}$ will be alternatively denoted $p \curvearrowright q$, and $p \curvearrowright q \oplus q \curvearrowright r$ will denote the broken geodesic [1] connecting p to r with a "stopover" at q. A probability distribution $\mathfrak{L}^i$ will be identified with its coordinates with respect to some chosen parametrization; for instance, we will write $\mathfrak{L}^i \equiv \left(\phi^i, \mu^i\right)$.

We also consider for any $x \in \mathfrak{M}$ the local norm $\|V\|_g(x)$ associated with the metrics g on the tangent space $T_x\mathfrak{M}$:

$$\forall V \in T_x\mathfrak{M},\ \|V\|_g(x) := \sqrt{V'.g(x).V}. \tag{1}$$

Finally, the length of a curve α traced on $\mathfrak{M}$ will be denoted $\Lambda(\alpha)$.

3 The Rao's distance

In a seminal paper, Rao [11] noticed that, equipped with the Fisher information metrics denoted $\mathfrak{g}(\bullet)$, a family of probabilities depending on p parameters can be considered as a p-dimensional Riemannian manifold. The associated Riemannian (Rao's) distance between the distributions with parameters $\theta^{(1)}$ and $\theta^{(2)}$ is given by:

$$D_{\mathcal{R}}\left(\theta^{(1)}, \theta^{(2)}\right) := \int_0^1 \sqrt{\dot{\gamma}'(t).\mathfrak{g}(\gamma(t)).\dot{\gamma}(t)}dt \tag{2}$$

where γ is a **segment** (minimal length curve) connecting $\theta^{(1)} = \gamma(0)$ to $\theta^{(2)} = \gamma(1)$. As any Riemannian distance, $D_{\mathcal{R}}$ is **intrinsic** (*i.e.* it is coordinates-free).

Riemannian geometry in a nutshell

Definition 3.1.
[1] Consider the differentiable manifold $\mathfrak{M}$, and the set $\mathcal{X}(\mathfrak{M})$ of vector fields on $\mathfrak{M}$. A linear connection (or covariant derivative) $\boldsymbol{D}$ on $\mathfrak{M}$ is a bilinear map

$$\begin{cases} \boldsymbol{D} : \mathcal{X}(\mathfrak{M}) \times \mathcal{X}(\mathfrak{M}) \to \mathcal{X}(\mathfrak{M}) \\ (X, Y) \mapsto \boldsymbol{D}_X Y \end{cases}$$

which is linear in X and a derivation on Y.

According to the fundamental theorem of Riemannian geometry [1], there is a unique symmetric connection ∇ compatible with a fixed metrics $\mathfrak{g}$ (the so-called Levi-Civita or Riemann connection), giving in our case the Rao's distance.

Definition 3.2.
[1, 5] Let $\gamma : [0,1] \to \mathfrak{M}$ be a curve traced on $\mathfrak{M}$, and $\boldsymbol{D}$ be a connection on $\mathfrak{M}$. γ is a geodesic with respect to $\boldsymbol{D}$ if its acceleration $\boldsymbol{D}_{\dot\gamma(t)}\dot\gamma(t)$ is null $\forall t \in]0,1[$. In other words, ***a geodesic has constant speed*** *in the local norm (1):*

$$\|\dot\gamma\|_{\mathfrak{g}} := \|\dot\gamma(\bullet)\|_{\mathfrak{g}}(\gamma(\bullet)) = \sqrt{\dot\gamma'(\bullet).\mathfrak{g}(\gamma(\bullet)).\dot\gamma(\bullet)}.$$

Corollary 3.1.
Let $\gamma : [0,1] \to \mathfrak{M}$ be a geodesic, and $[a,b] \subseteq [0,1]$. Then

$$\int_a^b \sqrt{\dot\gamma'(t).\mathfrak{g}(\gamma(t)).\dot\gamma(t)}dt = (b-a)\ \|\dot\gamma\|_{\mathfrak{g}}.$$

Geodesics on a p-dimensional Riemannian manifold with respect to ∇ are solutions of the Euler-Lagrange equation [5, 1, 3]:

$$\forall\, 1 \le k \le p,\ \ddot\gamma_k(t) + \sum_{i,j=1}^{p} \Gamma_{i,j}^k\, \dot\gamma_i(t)\, \dot\gamma_j(t) = 0 \tag{3}$$

where each coefficient of ∇ (some "Christoffel symbol" $\Gamma_{i,j}^k$) only depends on $\mathfrak{g}$, and is defined in coordinates by:

$$\Gamma_{i,j}^k := \sum_{m=1}^{p} \frac{\mathfrak{g}^{\mathrm{im}}}{2}\left(\frac{\partial \mathfrak{g}_{mj}}{\partial\theta_k} + \frac{\partial \mathfrak{g}_{mk}}{\partial\theta_j} - \frac{\partial \mathfrak{g}_{jk}}{\partial\theta_m}\right) \tag{4}$$

where $\mathfrak{g}^{\mathrm{im}}$ (resp. $\mathfrak{g}_{mk}$) is some entry of $\mathfrak{g}^{-1}$ (resp. $\mathfrak{g}$).

To determine the shortest curve between two points of $\mathfrak{M}$, one applies the following result.

Lemma 3.1.
[5, 1] Let $\mathfrak{M}$ be an abstract surface, and $p, q \in \mathfrak{M}$. Suppose that $\alpha : [a,b] \to \mathfrak{M}$ is a curve of minimal length connecting p to q. Then, α is a geodesic.

Nevertheless, building the segment connecting p to q is not straightforward, since the lemma above only shows that a segment is a geodesic. But a geodesic is not necessarily a segment...

Theorem 3.1.
[1] Let $p = \alpha(0)$ be the initial point of a geodesic. Then there is some $0 < t_0 \leq +\infty$ such that α is a segment from p to $\alpha(t)$ for every $t \leq t_0$ and for $t > t_0$ thereafter never again a segment from p to any $\alpha(t)$ for $t > t_0$. This number t_0 is called the cut value of α and $\alpha(t_0)$ is called the cut point of α. There are only two possible reasons (which can occur simultaneously) for $\alpha(t_0)$ to be to be the cut point of α:

- there is a segment from p to $\alpha(t_0)$ different from α

- $\alpha(t_0)$ is the first conjugate point on α to p (i.e. $t_0\,\dot{\alpha}(0)$ is a critical point of the exponential map, defined hereunder).

Remark 3.1.
No matter the cause of the phenomenon, the main point for us is that if t_0 is a cut value of α, $\forall\, t \leq t_0$, $D_{\mathcal{R}}(p, \alpha(t)) = t$ while $\forall\, t > t_0$, $D_{\mathcal{R}}(\alpha(t_0), \alpha(t)) < t - t_0$.

Definition 3.3.
[1] Let $\mathfrak{M}$ be a Riemann manifold and $x \in \mathfrak{M}$. The exponential map of $\mathfrak{M}$ at x is $\exp_x : W_x \to \mathfrak{M}$, defined on some neighborhood W_x of 0 in the tangent space $T_x\mathfrak{M}$ by:

$$\exp_x(V) := \alpha_{\mathcal{B}(V)}(\|V\|)$$

where $\mathcal{B}(V)$ is the projection of V onto the unit ball and $\alpha_{\mathcal{B}(V)}$ is the unique geodesic in $\mathfrak{M}$ such that $\alpha_{\mathcal{B}(V)}(0) = x$ and $\dot{\alpha}_{\mathcal{B}(V)}(0) = \mathcal{B}(V)$.

Remark 3.2.
If $\alpha := p \curvearrowright q$ is a segment and $V_0 := \dot{\alpha}(0)$, because of uniqueness of geodesics, $\exp_p(V_0) := \alpha_{\mathcal{B}(V_0)}(1) = q$; reciprocally, if $V_1 := \dot{\alpha}(1)$, $\exp_q(V_1) := \alpha_{\mathcal{B}(V_1)}(1) = p$ (compare Figures 1 & 2).

4 The geometry of $NB(D_{\mathcal{R}})$

The most classical parametrization of the NB distribution is given by

$$P(X = j; (\phi, p)) = \left(\begin{array}{c} \phi + j - 1 \\ j \end{array} \right) p^j (1-p)^{\phi} \; j \geq 0 \tag{5}$$

with $(\phi, p) \in \mathbb{R}^+ \times]0,1[$; ϕ is the index parameter (denoted k by [2] and many other authors). Nevertheless, because of its orthogonality, we chose instead the parametrization used by Chua and Ong [4]:

$$P(X = j; (\phi, \mu)) = \left(\begin{array}{c} \phi + j - 1 \\ j \end{array} \right) \left(\frac{\mu}{\mu + \phi} \right)^j \left(1 - \frac{\mu}{\mu + \phi} \right)^{\phi}, \; j \geq 0 \tag{6}$$

$(\phi, \mu) \in \mathbb{R}^+ \times \mathbb{R}^+$; here, μ is the mean of the distribution. In these coordinates, the information matrix is:

$$\mathfrak{g}(\phi, \mu) = \left(\begin{array}{cc} G_{\phi\phi} & 0 \\ 0 & G_{\mu\mu} \end{array} \right)$$

with $G_{\mu\mu} = \frac{\phi}{\mu(\mu+\phi)}$, while the expression of $G_{\phi\phi}$ is more complicated:

$$G_{\phi\phi} = -\frac{\mu + \phi(\mu + \phi)\left((\phi/\mu+\phi)^{\phi} - 1 \right)\psi^1(\phi)}{\phi(\mu + \phi)} \tag{7}$$

where ψ^1 is the Trigamma function (first derivative of the logarithmic derivative of $\Gamma(\bullet)$). One will find in [3] the closed-form expression of the Rao's distance for a number of probability families. These authors reported that when the index parameter of two NB distributions **is the same** the Rao's distance is given, in the parametrization (5), by:

$$D_{NB(p)}\left(\left(\phi,p^1\right),\left(\phi,p^2\right)\right) := 2\sqrt{\phi}\,\arccos\left(\frac{1-\sqrt{p^1\,p^2}}{\sqrt{\left(1-p^1\right)\left(1-p^2\right)}}\right). \tag{8}$$

Of course, if μ^1 (resp. μ^2) is the mean of $\mathfrak{L}^1 = NB\left(\phi,p^1\right)$ (resp. $\mathfrak{L}^2 = NB\left(\phi,p^2\right)$), we have necessarily:

$$D_{\mathcal{R}}\left(\mathfrak{L}^1,\mathfrak{L}^2\right) \leq D_{NB(p)}\left(\mathfrak{L}^1,\mathfrak{L}^2\right). \tag{9}$$

Due to the complexity of (7), $D_{\mathcal{R}}\left(\mathfrak{L}^1,\mathfrak{L}^2\right)$ cannot be obtained in a closed-form. It must be computed by finding the numerical solution of the Euler-Lagrange equation (3), completed in the parametrization (6) by the boundary conditions

$$\left\{\gamma\left(0\right)=\left(\phi^1,\mu^1\right),\,\gamma\left(1\right)=\left(\phi^2,\mu^2\right)\right\}. \tag{10}$$

Geodesics can be as well be computed by solving (3) under the alternative constraints

$$\left\{\gamma\left(0\right)=\left(\phi^1,\mu^1\right),\,\dot{\gamma}\left(0\right)=V\in\mathbb{R}^2\right\}. \tag{11}$$

This solution is associated with the exponential map at $\left(\phi^1,\mu^1\right)$.

5 Approximating $D_{\mathcal{R}}\left(\mathfrak{L}^1,\mathfrak{L}^2\right)$

In this section, $\mathfrak{L}^i \equiv \left(\phi^i,\mu^i\right)$ denotes a NB distribution parametrized in the (6) system, but our purpose could be extended to any parametric family.

Firstly, all the Christoffel symbols (4) were calculated from the expression (7) of $G_{\phi\phi}$, with the help of *Mathematica*. Then, the differential equation (3) was numerically solved under the the boundary conditions (10), for the estimated parameters of a number of marine organisms. In most case a solution could be found in an acceptable time (four CPU minutes, at most), with a good numerical precision (15 digits), but was each one of the geodesics found a segment? And what about failures in computation? We indeed had to face two different problems: a theoretical one and a computational one.

Theoretical issue

Suppose a solution $\gamma = \mathfrak{L}^1 \curvearrowright \mathfrak{L}^2$ of (3) under the boundary condition (10) has been found; according to Corollary 3.1, a straightforward approximation of $D_{\mathcal{R}}\left(\mathfrak{L}^1,\mathfrak{L}^2\right)$ should be $\|\dot{\gamma}\|_{\mathfrak{g}}$. But notice that $\|\dot{\gamma}\|_{\mathfrak{g}}$ is only an **upper bound**, which is attained only when there is **no cut point** in $\gamma\left([0,1]\right)$ (cf. Theorem 3.1). That is why we need some test to detect a possible cut point on some geodesic curve (see Section 5). Suppose now a cut point $\left(\phi^{c(1,2)},\mu^{c(1,2)}\right)$ has been detected on γ. Then, it is natural [1] to supersede γ by the broken geodesic

$$\left(\phi^1,\mu^1\right) \curvearrowright \left(\phi^{c(1,2)},\mu^{c(1,2)}\right) \oplus \left(\phi^{c(1,2)},\mu^{c(1,2)}\right) \curvearrowright \left(\phi^2,\mu^2\right)$$

whose length is shorter than $\Lambda\left(\gamma\right)$, provided $\left(\phi^{c(1,2)},\mu^{c(1,2)}\right) \curvearrowright \left(\phi^2,\mu^2\right)$ is also a segment.

Computational issues

Wet met various numerical problems in computing $\mathfrak{L}^1 \curvearrowright \mathfrak{L}^2$:

(P1) no solution was found (due to time limit, singularities, *etc*)

(P2) an unsuitable solution was found: for some $t \in [0,1]$, $(\phi(t), \mu(t)) \notin \mathbb{R}^+ \times \mathbb{R}^+$

(P3) the boundary condition (10) was not fulfilled with a satisfactory precision.

Simple configurations

When none of these issues is met, we first check that there is no cut point on γ. Then, the canonical solution is acceptable, and we can write:

$$D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right) \approx \Lambda(\gamma) = \|\dot{\gamma}\|_{\mathfrak{g}}. \tag{12}$$

If a cut point $\left(\phi^{c(1,2)}, \mu^{c(1,2)}\right)$ is detected on γ, and if $\left(\phi^{c(1,2)}, \mu^{c(1,2)}\right) \curvearrowright \left(\phi^2, \mu^2\right)$ is free of cut point, we adopt as an upper bound for $D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right)$:

$$\Lambda\left(\left(\phi^1, \mu^1\right) \curvearrowright \left(\phi^{c(1,2)}, \mu^{c(1,2)}\right)\right) + \Lambda\left(\left(\phi^{c(1,2)}, \mu^{c(1,2)}\right) \curvearrowright \left(\phi^2, \mu^2\right)\right).$$

Intricate configurations

When (P1) or (P2) is met, we consider that the best achievable solution would consist in breaking $\gamma = \mathfrak{L}^1 \curvearrowright \mathfrak{L}^2$ by inserting a well-placed "stopover". But since γ is undetermined, how should $\left(\phi^{S(1,2)}, \mu^{S(1,2)}\right)$ be chosen? We propose two heuristics for approaching γ:

1. compute a "rough solution" $\widetilde{\gamma_R}$ to the original problem, contenting ourselves with low-precision (here: 5 digits), and substitute $\widetilde{\gamma_R}$ for γ to search for $\left(\phi^{S(1,2)}, \mu^{S(1,2)}\right)$
2. when $\widetilde{\gamma_R}$ cannot be obtained, merely use instead $\widetilde{\gamma_L}(t) := t\left(\phi^1, \mu^1\right) + (1-t)\left(\phi^2, \mu^2\right)$.

In the second case, after fixing a convenient sampling rate $\frac{1}{N}$, the stopover naturally corresponds to the shortest broken geodesic:

$$\begin{cases} \left(\phi^{S(1,2)}, \mu^{S(1,2)}\right) = \widetilde{\gamma_L}\left(\frac{k_L}{N}\right) \\ k_L := \arg \min\limits_{1 \le k \le N-1} \left(\Lambda\left(\left(\phi^1, \mu^1\right) \curvearrowright \widetilde{\gamma_L}\left(\frac{k}{N}\right)\right) + \Lambda\left(\widetilde{\gamma_L}\left(\frac{k}{N}\right) \curvearrowright \left(\phi^2, \mu^2\right)\right)\right). \end{cases} \tag{13}$$

In the first case two eventualities must be considered:

1. a cut point $\left(\phi^{c(1,2)}, \mu^{c(1,2)}\right)$ is detected on $\widetilde{\gamma_R}([0,1])$; then $\left(\phi^{S(1,2)}, \mu^{S(1,2)}\right) = \left(\phi^{c(1,2)}, \mu^{c(1,2)}\right)$
2. if no cut point is detected, proceed like in (13):

$$\begin{cases} \left(\phi^{S(1,2)}, \mu^{S(1,2)}\right) = \widetilde{\gamma_R}\left(\frac{k_R}{N}\right) \\ k_R := \arg \min\limits_{1 \le k \le N-1} \left(\Lambda\left(\left(\phi^1, \mu^1\right) \curvearrowright \widetilde{\gamma_R}\left(\frac{k}{N}\right)\right) + \Lambda\left(\widetilde{\gamma_R}\left(\frac{k}{N}\right) \curvearrowright \left(\phi^2, \mu^2\right)\right)\right). \end{cases} \tag{14}$$

Boundary problems

(P3) is easy to solve, since it merely corresponds to $\gamma(0) \neq \mathfrak{L}^1$ or $\gamma(1) \neq \mathfrak{L}^2$. We just have to add to formulas (12), (13) or (14) the corrective boundary error term

$$BE(\gamma) := \left\|\gamma(0) - \mathfrak{L}^1\right\|_{\mathfrak{g}} \left(\mathfrak{L}^1\right) + \left\|\gamma(1) - \mathfrak{L}^2\right\|_{\mathfrak{g}} \left(\mathfrak{L}^2\right) \tag{15}$$

given by formula (1). Finally, we can write:

$$D_{\mathcal{R}}\left(\mathfrak{L}^1, \mathfrak{L}^2\right) \leq \Lambda(\gamma) + BE(\gamma) \tag{16}$$

whatever the selected geodesic (broken, or not) may be.

Locating a (N, ϵ)- cut point on some geodesic γ

For that purpose, the unit interval is first divided into N intervals: $[0,1] = \bigcup_{i=1}^{N} \delta_i$, with $\delta_i := [\frac{i-1}{N}, \frac{i}{N}[$. Suppose there exists a cut point $\gamma(t_c)$ on γ, such that $t_c \in \delta_{i_c}$. Consider the set

$$\mathfrak{C}_N(\gamma) := \left\{\gamma_1 := \gamma\left(\frac{1}{N}\right), \cdots, \gamma_k := \gamma\left(\frac{k}{N}\right), \cdots, \gamma_{N-1} := \gamma\left(\frac{N-1}{N}\right)\right\} \subset \mathfrak{M}$$

and, for each $1 \leq i \leq N$ the geodesic $\alpha_i := \gamma_{i-1} \curvearrowright \gamma_i$ obtained by solving (3) **under the constraints**

$$\{\alpha_i(0) = \gamma_{i-1},\ \alpha_i(1) = \gamma_i\}.$$

Because of the uniqueness of segments, Corollary 3.1 and Remark 3.1, $\forall\, i < i_c$, $\frac{\|\dot{\gamma}\|_{\mathfrak{g}}}{N} = \Lambda(\alpha_i) = \|\dot{\alpha_i}\|_{\mathfrak{g}}$. On the contrary, when $i \geq i_c$, the distance between γ_{i-1} and γ_i **along** γ is $\frac{\|\dot{\gamma}\|_{\mathfrak{g}}}{N}$ yet, while $\|\dot{\alpha_i}\|_{\mathfrak{g}}$ should be smaller. More precisely, if the resolution $\frac{1}{N}$ is small enough (for instance, smaller than the injectivity radius [1] of $\mathfrak{M}$), $\gamma_{i-1} \curvearrowright \gamma_i$ is a segment and we may write:

$$\begin{cases} \forall\, i < i_c,\ \frac{\|\dot{\gamma}\|_{\mathfrak{g}}}{N} - \|\dot{\alpha_i}\|_{\mathfrak{g}} = 0 \\ \forall\, i \geq i_c,\ \frac{\|\dot{\gamma}\|_{\mathfrak{g}}}{N} - \|\dot{\alpha_i}\|_{\mathfrak{g}} > 0. \end{cases}$$

Thus, after fixing ϵ (small), we can locate possible cut points, with a precision depending on (N, ϵ).

Definition 5.1.
We will say that $\gamma_{i_c} \in \mathfrak{C}_N(\gamma)$ is a (N, ϵ)- cut point on γ if

$$i_c = \arg \min_{1 \leq i \leq N-1} \left(\left| \frac{\|\dot{\gamma}\|_{\mathfrak{g}}}{N} - \|\dot{\alpha_i}\|_{\mathfrak{g}} \right| > \epsilon \right).$$

6 The MEEZ data

The Mauritanian coast, situated on the Atlantic side of the northwestern African continent, embeds a wide long continental shelf of about $750\,km$ and $36000\,km^2$ with an Exclusive Economic Zone (MEEZ) of 230000 km^2. This study focuses on the analysis of abundance of fish and invertebrates data collected during annual scientific trawl surveys performed by oceanographic vessels on the continental shelf ($<$ 200 m depth), from 1987 to 2010. All the species (fish and

invertebrate) captured in a given station were identified, counted and then recorded on the database. In addition, each station has been characterized by supplementary environmental variables: bathymetry, sedimentary type of the substrate, latitude and longitude. The counts of species collected were then fitted by NB distributions. For that purpose, it was necessary to determine homogeneous regions (habitats) in the MEEZ; it was found that the optimal number of habitats is four. Then the counts of each species were separately fitted in each one of these regions, and it was observed that in each one of the habitats, only a reduced number of species could be satisfactorily fitted by some NB distribution; other species were discarded. For further details on the data or estimation methods, see [9, 7].

7 Results

Geodesics: a bestiary

We illustrate hereunder the diversity of cases encountered in computing $D_{\mathcal{R}}(A, B)$. From now, the approximation parameter are fixed to $(N, \epsilon) = (10, 0.01)$. All the figures displayed will be composed of three panels. On the left one, we superimposed the final solution to the rough geodesic (when it could be computed). On both the other panels, we investigated the structure of broken geodesics in the neighborhood of a stopover S, with the help of the exponential map. We determined first $\gamma_1 = A \curvearrowright S$ (resp. $\gamma_2 = B \curvearrowright S$) by solving equation (3) under the constraints (10). We afterward considered $\{V_i(\theta_k) := \rho(\theta_k) \centerdot \mathcal{B}(\dot{\gamma}_i(0)) : i = 1, 2\}$, where the angle of the rotation ρ is (in degrees) $\theta_k \in \{0, \pm 0.1, \pm 0.2, \pm 0.3\}$. Equation (3) was then solved under the constraints (11) with $V = V_i(\theta_k)$, giving rise to two bundles of seven geodesics; remember that for $\theta = 0$, $\exp_A(V_1(0)) = S = \exp_B(V_2(0))$ (see Remark 3.2). In all these plots, the red point will be "A" and the black one will be "B", while the stopover is represented by the big gray point.

On Figure 1, we represented the geodesic $\gamma_1 := A \curvearrowright B$, with $A \equiv (0.7767, 11.2078)$ and $B \equiv (0.7767, 87.268)$ in the system (6). It corresponds to a simple configuration: no (N, ϵ)- cut point was found, and we can see on the left panel that there is practically no difference between the segment and the sampled rough geodesic. We stress that the stopover S is in this case quite unnecessary; it was introduced only for illustration. On the central panel, the segment $A \curvearrowright S$ has been extrapolated with the exponential map, as well as the other geodesics of the bundle. On the right panel the segment $B \curvearrowright S$ and the corresponding bundle of geodesics have been extrapolated in the same way. We can see that there is practically no difference between extrapolations of $A \curvearrowright S$ and $B \curvearrowright S$, the segment γ_1 and the rough geodesic $\widetilde{\gamma_{1,R}}$. Notice finally that in this (artificial) case, the distance $D_{NB(p)}$ (8) given by [3] can be computed. In the (5) system, $A \equiv (0.7767, 0.935191)$, $B \equiv (0.7767, 0.991178)$ and we have, in compliance with (9):

$$1.7 \approx D_{\mathcal{R}}(A, B) < D_{NB(p)}(A, B) = 1.783.$$

On Figure 2, we plotted geodesics connecting two species: *Rhizoprionodon acutus*, coded RIAC70, and *Anguilla sp*, coded ANSP50. Even if $RIAC70 \curvearrowright ANSP50$ could not be determined, the rough geodesic $\widetilde{\gamma_R}$ could be computed. A (N, ϵ)- cut point was detected in the second position of the sampled curve, and used as a stopover S to compute the final broken segment. But we can see of the central and the right panels that neither $A \curvearrowright S$ nor $B \curvearrowright S$ could be extrapolated to obtain a geodesic connecting A to B.

Figure 1. A geodesic without any cut point $\left(\widetilde{D_{\mathcal{R}}} \approx 1.70 \approx D_{\mathcal{R}}\right)$. Left panel: the rough geodesic (cyan suits) is superimposed to the segment; A is the red point, B the black one and the stopover is represented by the gray point. Right panels: plot of the two bundles of geodesics issued from A or B. Red curve: $\theta = 0$; dashed curves: $\theta \neq 0$. The header corresponds to the parameters of the distributions.

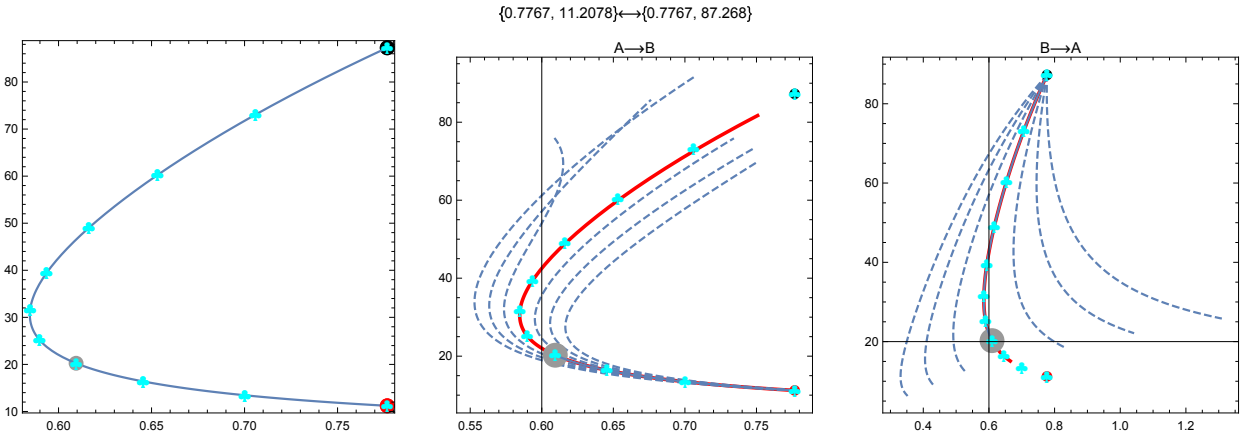


Figure 2. A geodesic with a cut point; $\widetilde{D_{\mathcal{R}}}\left(RIAC70, ANSP50\right) \approx 3.98$ and $D_{\mathcal{R}}\left(ANSP50, RIAC70\right) \approx 3.90$. Same graphical conventions as in Figure 1.

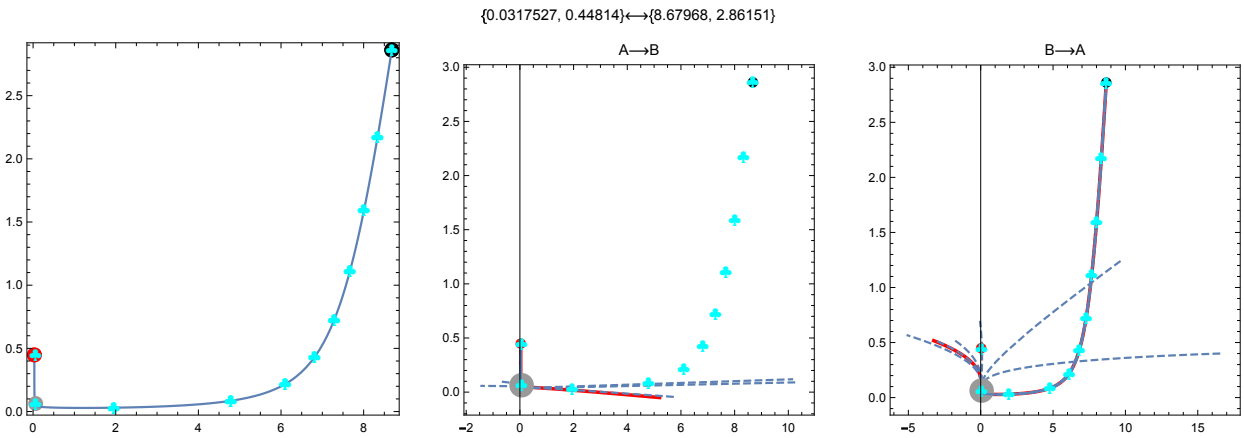


Figure 3. A broken geodesic between two species; $D_{\mathcal{R}}\left(HISP00, SCAN40\right) \approx 29.62$. Same graphical conventions as in Figure 1.

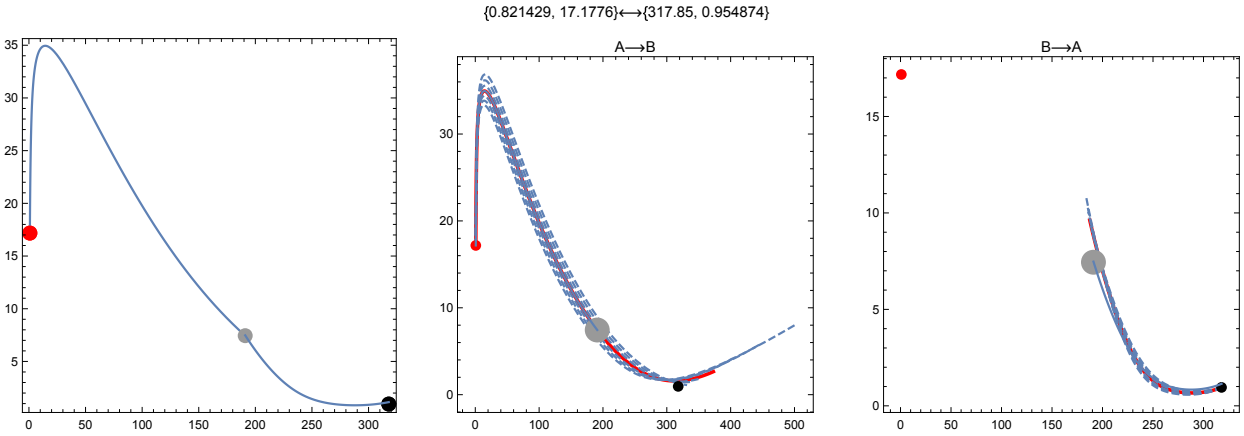
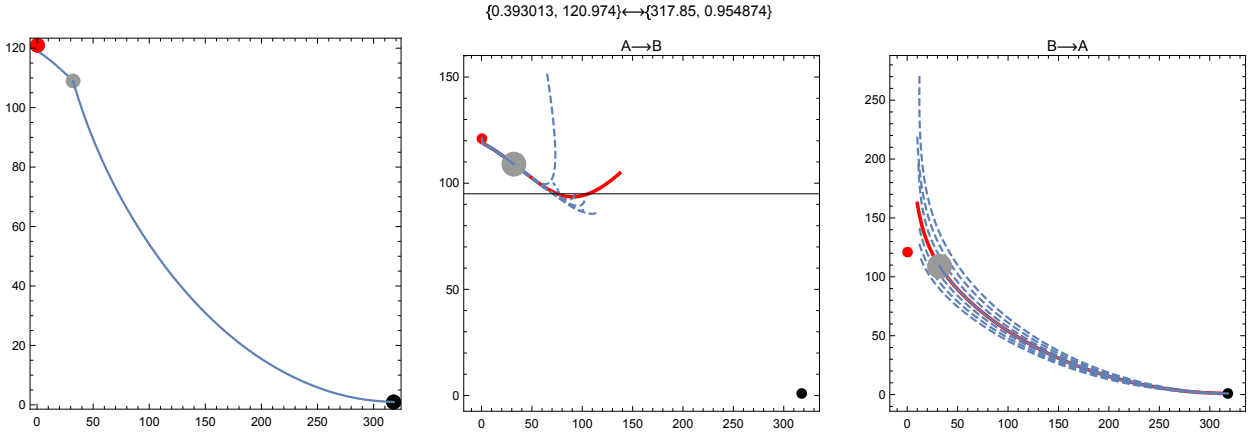


Figure 4. A broken geodesic between two species; $D_{\mathcal{R}}\left(HISP00, TRTR20\right) \approx 30.60$. Same graphical conventions as in Figure 1.



Now, what about the worst cases, when linear interpolation was unavoidable? There were 4 such pairs of species in the habitat $C4$ (see Table 1). Notice first that all these pairs were associated with a particular species, of parameters $(317.85, 0.954874)$: this is *Hippocampus sp*, coded HISP00, the less aggregative species in this habitat.

Let us start with $\{HISP00, SCAN40\}$, whose processing is represented on Figure 3 (SCAN40 is the code of *Scorpaena angolensis*). In this case, neither of the geodesics could be computed, and we used in last resort linear interpolation in the space of parameters. The obtained curve is rather smooth, and one could probably find a genuine segment close to this broken geodesic, with enough computation time.

Another example: $D_{\mathcal{R}}\left(HISP00, TRTR20\right)$, where TRTR20 is the code of *Trachurus trecae*. This case, displayed on Figure 4, is quite different: the structure of the geodesics near the stopover S looks like the structure of geodesics in the neighborhood of a cut point (see Figure 2). But notice S was found by traveling across $\widetilde{\gamma_L}\left(\bullet\right)$, and one cannot claim it is a realistic first guess for $HISP00 \curvearrowright TRTR20$.

Return to the exploratory setting

Remember that the MEEZ could be split into four homogeneous regions (see Section 6), named $\{C_1, \cdots, C_4\}$. From the estimates of the parameters of the N_h species kept in C_h, it is possible to tabulate the Rao's distance between species and process the resulting table with methods designed for non-Euclidean distances (Multidimensional Scaling, Isomap, *etc*), as proposed by Rao [11] himself. Because of the computational cost of Rao's distances, we were forced to select, for each habitat, a sub-sample of species representing as well as possible the whole (landmark species, say). Thus, in $C4$ (like in other habitats), species were first split into two categories: very aggregative and moderately aggregative. We focused on the second category, keeping for computation the 30-species set (amongst the 121 species correctly fitted, while 301 species were observed) obtained by gathering isolated species and species constituting the vertices of the convex envelope of non-isolated species (see Figure 6 of [9]).

Global statistics

It is interesting to tally the various configurations encountered in different habitats: simple or intricate, and the presence of possible (N, ϵ)-cut points on the obtained geodesics. In the intricate case, it is also interesting to tally the cases where linear interpolation was unavoidable. The obtained results are gathered in Table 1. More than 70% of the configurations (88% in $C4$) were simple (*i.e.* the canonical solution was accepted), and (N, ϵ)- cut points were quite rare. In the intricate cases, the rough solution was generally accepted (more than 90% of the cases). We can thus claim that the obtained upper bounds given by Formula(16) were mostly tight approximations of true Rao's distance.

Table 1. Global results obtained in the four habitats of the MEEZ

Habitat	Number of species (well-fitted)	Simple configurations	Intricate (Rough, Linear)	Cut points
$C1$	30	356	(75,4)	1
$C2$	19	124	(46,1)	2
$C3$	26	227	(88,10)	1
$C4$	26	288	(33,4)	1

Acknowledgments

We thank the Mauritanian Institute of Oceanographic Research and Fisheries (IMROP) and the Department of Cooperation and Cultural Action of the Embassy of France in Mauritania for their support for this study. We also thank all scientists who contributed to field surveys and data collection.

Bibliography

[1] Berger, M. (2003) *A Panoramic View of Riemannian Geometry*. Springer Verlag, Berlin.

[2] Bliss, C. I. and Fisher, R. A. (1953) *Fitting the Negative Binomial distribution to biological data*. Biometrics, **9**, 176-200.

[3] Burbea, J. and Rao, C. R. (1986) *Infomative geometry of probality spaces*. Expo. Math., **4**, 347-378.

[4] Chua, K. C. and Ong, S. H. (2013). *Test of mispecification with application to Negative Binomial distribution*. Computational Statistics, **28**, 993-1009.

[5] Gray, A. (1999) *Modern differential geometry of curves and surfaces with Mathematica* (2nd ed.). CRC Press, London.

[6] Kendall, D.G. (1948). *On some modes of population growth leading to R. A. Fisher's logarithmic series distribution.* Biometrika, **35**, 6-15.

[7] Kidé, S. O., Manté, C., Dubroca, L., Demarcq, H., Mérigot, B. (2015) *Spatio-Temporal Dynamics of Exploited Groundfish Species Assemblages Faced to Environmental and Fishing Forcings: Insights from the Mauritanian Exclusive Economic Zone*. PLoSONE, **10, 10**, e0141566. doi:10.1371/journal.pone.0141566

[8] Manté, C., Durbec, J.P. and Dauvin, J. C. (2005) *A functional data-analytic approach to the classification of species according to their spatial dispersion. Application to a marine macrobenthic community from the Bay of Morlaix (western english channel)*. J. Appl. Statist., **32, 8**, 831-840.

[9] Manté, C., Kidé, O. S., Yao, A. F. and Mérigot, B. (2016) *Fitting the truncated negative binomial distribution to count data. A comparison of estimators, with an application to groundfishes from the Mauritanian Exclusive Economic Zone*. Environmental and Ecological Statistics, DOI 10.1007/s10651-016-0343-1.

[10] O'Neil, M. F. and Faddy, M. J. (2003) *Use of binary and truncated negative binomial modelling in the analysis of recreational catch data*. Fisheries Research, **60**, 471-477.

[11] Rao, C. R. (1992) *Information and accuracy attainable in the estimation of statistical parameters*. In: *Breakthroughs in Statistics: Foundations and Basic Theory*. Springer, New York, 235-247.

[12] Vaudor, L., Lamouroux, N. and Olivier, J. M. (2011) *Comparing distribution models for small samples of overdispersed counts of freshwater fish*. Acta Oecologica, **37, 3**, 170-178.